VIRUS EVOLUTION

*Virus Evolution*, 2022, **8(2)**, 1–11

DOI: https://doi.org/10.1093/ve/veac087
Advance access publication date: 16 September 2022
**Research Article**

# Host prediction for disease-associated gastrointestinal cressdnaviruses

Cormac M. Kinsella,[1,2,*,†] Martin Deijs,[1,2] Christin Becker,[3] Patricia Broekhuizen,[4] Tom van Gool,[4,2] Aldert Bart,[5,‡] Arne S. Schaefer,[3] and Lia van der Hoek[1,2,*]

[1]Amsterdam UMC, Laboratory of Experimental Virology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands, [2]Amsterdam Institute for Infection and Immunity, Postbus 22660, Amsterdam 1100 DD, The Netherlands, [3]Department of Periodontology, Oral Surgery and Oral Medicine, Institute for Dental and Craniofacial Sciences, Berlin Institute of Health, Charité—Universitätsmedizin Berlin, Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany, [4]Amsterdam UMC, Laboratory of Clinical Parasitology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands and [5]Department of Medical Microbiology, Tergooi MC, Van Riebeeckweg 212, Hilversum 1213 XZ, The Netherlands

[†]https://orcid.org/0000-0001-9865-4608

[‡]https://orcid.org/0000-0001-6605-6347

*Corresponding authors: E-mail: c.m.kinsella@amsterdamumc.nl; c.m.vanderhoek@amsterdamumc.nl

## Abstract

Metagenomic techniques have facilitated the discovery of thousands of viruses, yet because samples are often highly biodiverse, fundamental data on the specific cellular hosts are usually missing. Numerous gastrointestinal viruses linked to human or animal diseases are affected by this, preventing research into their medical or veterinary importance. Here, we developed a computational workflow for the prediction of viral hosts from complex metagenomic datasets. We applied it to seven lineages of gastrointestinal cressdnaviruses using 1,124 metagenomic datasets, predicting hosts of four lineages. The *Redondoviridae*, strongly associated to human gum disease (periodontitis), were predicted to infect *Entamoeba gingivalis*, an oral pathogen itself involved in periodontitis. The *Kirkoviridae*, originally linked to fatal equine disease, were predicted to infect a variety of parabasalid protists, including *Dientamoeba fragilis* in humans. Two viral lineages observed in human diarrhoeal disease (CRESSV1 and CRESSV19, i.e. pecoviruses and hudisaviruses) were predicted to infect *Blastocystis* spp. and *Endolimax nana* respectively, protists responsible for millions of annual human infections. Our prediction approach is adaptable to any virus lineage and requires neither training datasets nor host genome assemblies. Two host predictions (for the *Kirkoviridae* and CRESSV1 lineages) could be independently confirmed as virus–host relationships using endogenous viral elements identified inside host genomes, while a further prediction (for the *Redondoviridae*) was strongly supported as a virus–host relationship using a case–control screening experiment of human oral plaques.

**Key words:** cressdnavirus; host identification; protist; *Redondoviridae*; metagenomics; periodontitis.

## Introduction

A defining feature of viruses is their obligate relationship with hosts, yet surprisingly hosts of most newly identified viruses remain unknown (Simmonds et al. 2017; Dolja and Koonin 2018; Greninger 2018). This circumstance is driven by widespread use of high-throughput sequencing for the discovery of viral genomes (Shi et al. 2016; Tisza et al. 2020; Edgar et al. 2022) versus traditional techniques, such as viral isolation in cell culture. In particular, metagenomic sequencing of taxonomically diverse samples obscures virus–host relationships, because of the many potential pairings. Exemplifying this are the cressdnaviruses, a group with small circular ssDNA genomes encoding a replication-associated protein. Now classified under the phylum *Cressdnaviricota* (Krupovic et al. 2020), the vast majority have unknown hosts (Simmonds et al. 2017; Tisza et al. 2020). This even applies to notable disease-associated lineages identified frequently in the gastrointestinal tracts of humans and other animals, referred to hereafter as gastrointestinal cressdnaviruses (Li et al. 2015; Phan et al. 2016; Abbas et al. 2019; Ramos et al. 2021). Among these are the family *Redondoviridae*, residents of the human mouth and lung linked to both periodontitis and critical illness (Abbas et al. 2019; Zhang et al. 2021), and the *Kirkoviridae*, found variously in dead and diseased horses, cows, and pigs, and also in human stool (Shan et al. 2011; Li et al. 2015; Zhao et al. 2017; Guo et al. 2018; Xie et al. 2020). Because infectious gastrointestinal disease is a leading cause of global mortality and morbidity in humans and livestock (Tam et al. 2012; Kirk et al. 2015; Thumbi et al. 2015), there is a clear need to determine the hosts of gastrointestinal cressdnaviruses, data that will underpin their medical or veterinary relevance.

Historically, no host inference methodology was required for cressdnaviruses, since observation of host disease preceded

discovery of the responsible virus. For example, banana bunchy top disease was recognised from approximately 1880 and classified as viral in the 1920s (Magee 1927), before the responsible cressdnavirus of family *Nanoviridae* was characterised later in the 20th century (Harding et al. 1993). Similarly, plant diseases have been linked to the *Geminiviridae* (Varma and Malathi 2003), avian and porcine diseases to the *Circoviridae* (Ritchie et al. 1989; Ellis et al. 1998), fungal debilitation to the *Genomoviridae* (Yu et al. 2010), and diatom lysis to the *Bacilladnaviridae* (Nagasaki et al. 2005). The challenge of the metagenomic age will be the identification of hosts when only the viral genome is known. While promising wet-lab methods, such as single-cell sequencing (Yoon et al. 2011) or proximity ligation (Bickhart et al. 2019; Ignacio-Espinoza et al. 2020) will enable simultaneous virus discovery and linkage to host sequences in future, these techniques are still emerging in the viral metagenomics field, and offer no solution for the thousands of conventionally sequenced viruses. Here, high-throughput computational approaches are required. Indirect solutions such as genome compositional analyses (Kapoor et al. 2010) or machine learning have been suggested; however, the former requires host genome assemblies or validated training datasets and suffers from relatively low accuracy (Ahlgren et al. 2017; Liu et al. 2019), while the latter generally requires validated training datasets, making it most appropriate for host prediction within otherwise well-characterised lineages (Eng, Tong, and Tan 2014; Babayan, Orton, and Streicker 2018).

Viral fossils inside host genomes, such as CRISPR spacers or endogenous viral elements (EVEs), provide direct evidence of virus–host relationships (Liu et al. 2011; Dion et al. 2021; Zhao, Lavington, and Duffy 2021). Among cressdnaviruses, the *Smacoviridae* have been proposed to infect archaea on the basis of matched CRISPR spacers (Díez-Villaseñor and Rodriguez-Valera 2019), while three families (*Naryaviridae, Nenyaviridae*, and *Vilyaviridae*) were linked to gut parasites using EVE evidence (Kinsella et al. 2020). Again, however, availability of host genome assemblies limits the range of this approach, and many virus–host relationships are likely unrepresented in the genomic fossil record. To date, no EVEs have been found belonging to the aforementioned redondoviruses or kirkoviruses. For such groups, high-throughput host prediction approaches that do not rely on host assemblies are needed. Here, we developed an analysis workflow for host prediction from metagenomic sequencing datasets, aiming to identify over-represented eukaryotes among virus positive samples for subsequent investigation. Through the analysis of 1,124 gastrointestinal tract samples, we could identify multiple cressdnavirus–eukaryote associations. Host predictions included redondoviruses with the human oral parasite *Entamoeba gingivalis*, kirkoviruses with parabasalid protists including *Dientamoeba fragilis* in humans, the CRESSV1 lineage (i.e. pecoviruses) with *Blastocystis* spp., and the CRESSV19 lineage (i.e. hudisaviruses) with *Endolimax nana*. Subsequent independent analysis confirmed several of these predictions as virus–host relationships.

# Results

## Census of gastrointestinal cressdnavirus lineages

Here, we aimed to predict the host of any cressdnavirus lineage displaying an apparently obligate association to the gastrointestinal tracts of vertebrates. Because no study has so far focused collectively on gastrointestinal cressdnaviruses, we first comprehensively censused published cressdnavirus sequences to determine the lineages meeting that definition. Iterative searches of the GenBank protein database collected 15,815 unique cressdnavirus Rep sequences, 2,461 of which remained after clustering. Each taxonomic class (*Arfiviricetes* and *Repensiviricetes*) was phylogenetically analysed separately (1,850 and 611 sequences, respectively). To work with unclassified lineages, clusters of related sequences were assigned a temporary name according to their branch support. We followed the format introduced by Kazlauskas, Varsani, and Krupovic (2018) who named the unclassified lineages CRESSV1 to CRESSV6. We added CRESSV7 to CRESSV33 in the *Arfiviricetes* (Fig. 1) and CRESSV34 to CRESSV39 in the *Repensiviricetes* (Supplementary Fig. S1). All previously named families and lineages were supported by our analysis, with the exception of CRESSV2, whose members remained adjacent but with poor branch support (Supplementary Fig. S2). We suggest that CRESSV2 may be most accurately characterised as multiple distinct lineages, here denoted as CRESSV2.1 to CRESSV2.6. Supporting this, the resulting sublineages showed unique isolation source patterns; for example, most members of CRESSV2.2 came from marine animal tissues and seawater, CRESSV2.3 were found predominately in human or livestock stool and tissue, and CRESSV2.4 members were identified in spiders, insects, and bird anal swabs (Supplementary Table S1).

Of fifty-six named lineages across the *Cressdnaviricota*, we categorised thirteen as putatively gastrointestinal due to their isolation source patterns (see Materials and methods). All were in the *Arfiviricetes* (Fig. 1, Supplementary Table S1). Four of these were excluded immediately because host inferences were already published; these were the *Smacoviridae, Naryaviridae, Nenyaviridae*, and *Vilyaviridae* (Díez-Villaseñor and Rodriguez-Valera 2019; Kinsella et al. 2020). Seven of the nine remaining lineages were found mainly in oral, gastrointestinal, or stool samples of various vertebrates, and some wastewater samples. These were the *Redondoviridae, Kirkoviridae*, CRESSV1 (i.e. pecoviruses), CRESSV2.3, CRESSV2.6, CRESSV9, and CRESSV19 (i.e. hudisaviruses). The others (CRESSV16 and CRESSV20) were detected predominately in wastewater, and were included since this source is often stool contaminated. The retained lineages were widely distributed phylogenetically, although notably some neighboured each other. For example, the lineage CRESSV9 was a close relative of the *Redondoviridae*, and together they were related to the *Naryaviridae*, viruses of *Entamoeba* parasites. Meanwhile, CRESSV19 and CRESSV20 clustered together, CRESSV1 was related to the *Giardia*-infecting *Vilyaviridae* and *Kirkoviridae* was related to both the lineage CRESSV2.3 and the *Nenyaviridae*, the latter also infecting *Entamoeba*.

## Recombination events and viral distributions reveal host biases

We identified nine gastrointestinal cressdnavirus lineages with unknown hosts. Some lineages were found in multiple vertebrate taxa, for example, CRESSV1 was known from stools of humans, pigs, and a camel, amongst others. This raised the possibility of different host preferences within a given lineage. Because this scenario would affect downstream analysis, we looked for viral recombination within lineages, which serves as evidence of shared host ranges since recombination must occur in the same host cell (Duffy, Burch, and Turner 2007; Kinsella et al. 2020). We first explored the genomic patterns of cressdnavirus recombination, analysing phylogenetic compatibility between nucleotide alignment windows along all available redondovirus genomes. This showed that highest incompatibility is found between genes, not within them (Fig. 2A), suggesting modular recombination of complete genes with different evolutionary histories. This pattern was

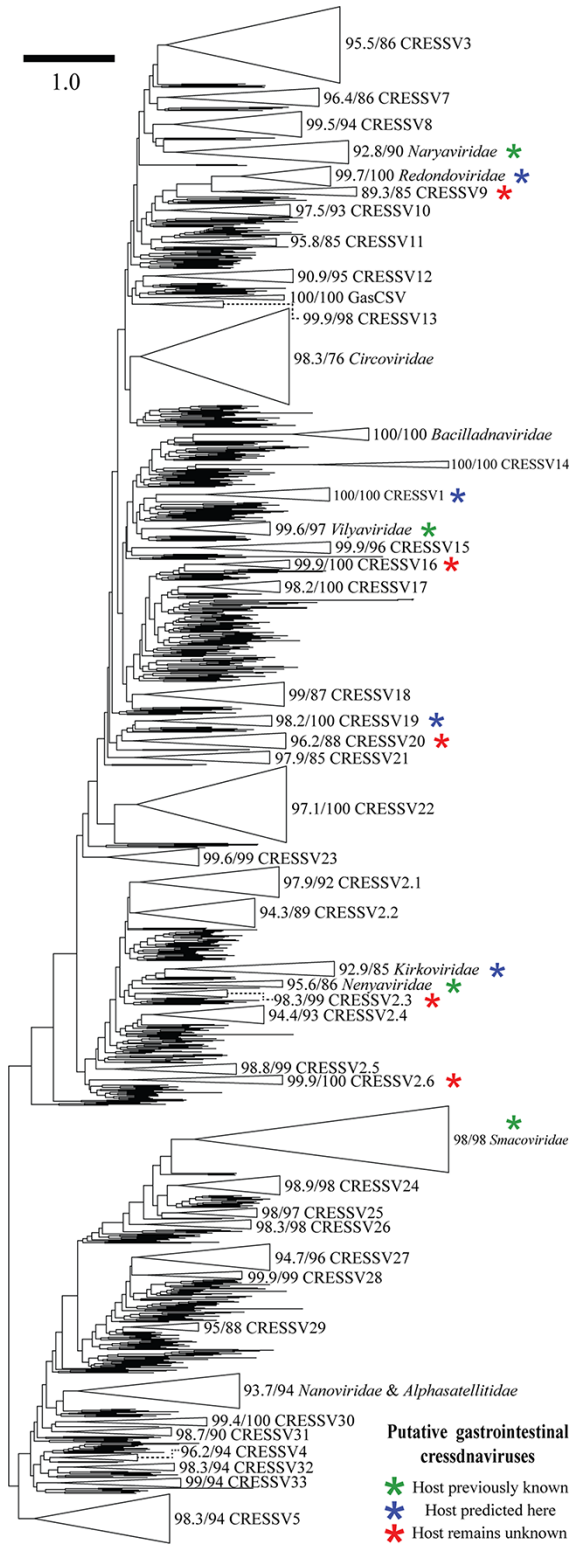


**Figure 1.** Maximum likelihood phylogenetic tree of the *Arfiviricetes*, rooted at the midpoint. Scale bar denotes amino acid substitutions per site. Branch supports are given for each named lineage, with SH-aLRT scores on the left and ultrafast bootstrap scores on the right. All sequences found outside of collapsed nodes did not meet criteria for naming a lineage.

corroborated by analysing the distribution of breakpoint pair coordinates, showing relative enrichment in two coordinate regions, those linking the start and end of the *Rep* gene, and those linking the start and end of the *Cap* gene (Fig. 2A). This propensity to swap genes as complete units is likely due to a reduced risk of protein structure disruption when compared with intra-gene recombination (Lefeuvre et al. 2007, 2009). We built on the observation by constructing tanglegrams between Rep and Cap protein phylogenies for each lineage (Fig. 2B–F, Supplementary Fig. S3). These provided some insight, for example, the extensive modular recombination among human-associated redondoviruses strongly suggests they share one host (Fig. 2B).

We annotated tanglegrams with reported isolation sources, finding that related viruses (sublineages) often shared sources. For example, pig-associated sublineages were observed in the *Redondoviridae* (Fig. 2B), the *Kirkoviridae* (Fig. 2C), CRESSV1 (Fig. 2D), and CRESSV9 (Fig. 2E). We hypothesised that such source biases might reflect varying host tropism, and to clarify this we used RDP4 to identify further recombination events within lineages. Interestingly, while most detected events occurred within sublineages, some gene flow was found between them (Fig. 2B–F). Overall, this suggested that members of each lineage overlapped in host range, yet displayed some specialisation at the sublineage level, perhaps to different host subtypes or species. An exception was a 'reproductively isolated' CRESSV1 sublineage found in birds (Fig. 2D), that displayed no evidence of recombination outside itself. To explicitly visualise source biases between human and porcine samples, we mapped the distribution of gastrointestinal cressdnaviruses across seven cohorts comprising 1,124 metagenomic sequencing datasets (Supplementary Tables S2 and S3). These were generated from human stool ($N = 374$), pig stool ($N = 512$), and human oral samples ($N = 238$). The analysis confirmed strongly biased distributions for some viruses, for example, members of CRESSV9 and *Kirkoviridae* were either strictly pig-associated or strictly human-associated across cohorts (Fig. 3). It also showed more flexible viruses, for example, some members of CRESSV1. Consistent with previous literature, we found that human-associated redondoviruses were the only lineage prevalent in the human oral environment, with more sporadic detection in stool (Abbas et al. 2019). Strikingly, previously unrecognised pig-associated redondoviruses (MT135242.1, KJ433989.1, and NC_035476.1) were entirely absent from human oral samples, but highly prevalent in porcine stool. The analysis also revealed that CRESSV16 (previously included for its occurrence in wastewater) was not found in any sample, leading to its exclusion from further analyses. From these analyses, we concluded that members of a viral lineage found in one isolation source (e.g. pig stool) likely shared the same host.

## Viral host prediction

To identify potential hosts of gastrointestinal cressdnaviruses, eukaryotic rRNA content of all 1,124 samples was classified, resulting in taxon lists at the genus level. Individually, for the six cohorts using Illumina deep sequencing, samples highly positive for each virus lineage were identified and compared to pinpoint prevalent eukaryotic taxa. Thus, shortlists of theoretically possible host candidates were generated for each virus lineage/cohort intersection (Supplementary Table S4). The low number of samples positive for lineage CRESSV2.6 in any cohort excluded it from the analysis at this point, leaving seven lineages (although human-associated and pig-associated redondoviruses were analysed separately). Next, host predictions were made by assessing the statistical associations between viruses and respective host candidates across all samples of all seven cohorts. Human oral cohorts contained only one cressdnavirus lineage,

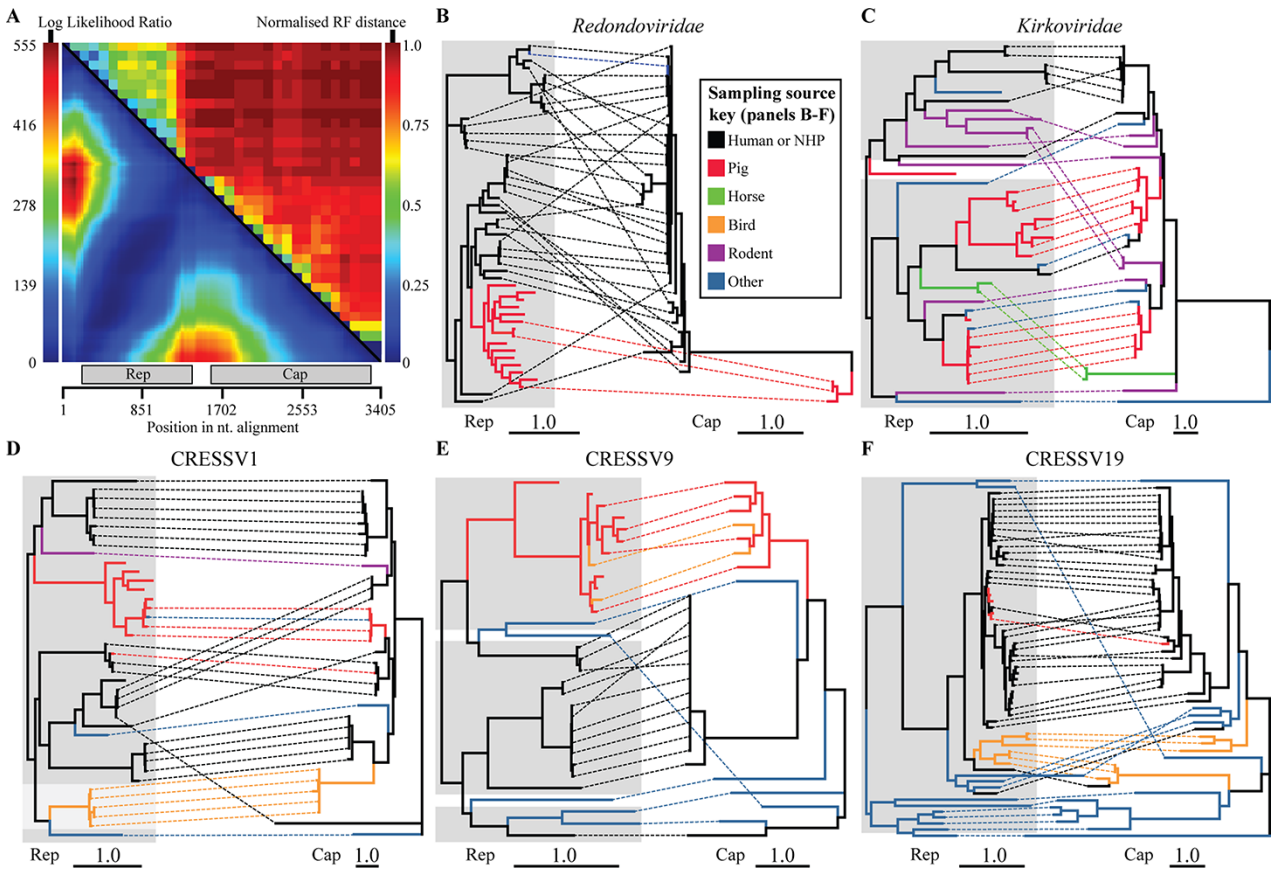


**Figure 2.** Recombination within gastrointestinal cressdnavirus lineages. (A) Upper right: phylogenetic compatibility matrix (Robinson-Foulds distance) computed on an alignment of redondovirus genomes, lower left: LARD breakpoint matrix computed on the same alignment. (B–F) Rep and Cap protein tanglegrams for five cressdnavirus lineages. Dotted lines connect proteins encoded by the same genome. Branch colour denotes isolation source as listed in the key. Grey blocks denote groups linked by RDP4 detected recombination events, and different shades represent different recombination groups (Panel D only). Scale bars on individual phylograms are in amino acid substitutions per site. NHP: non-human primate.

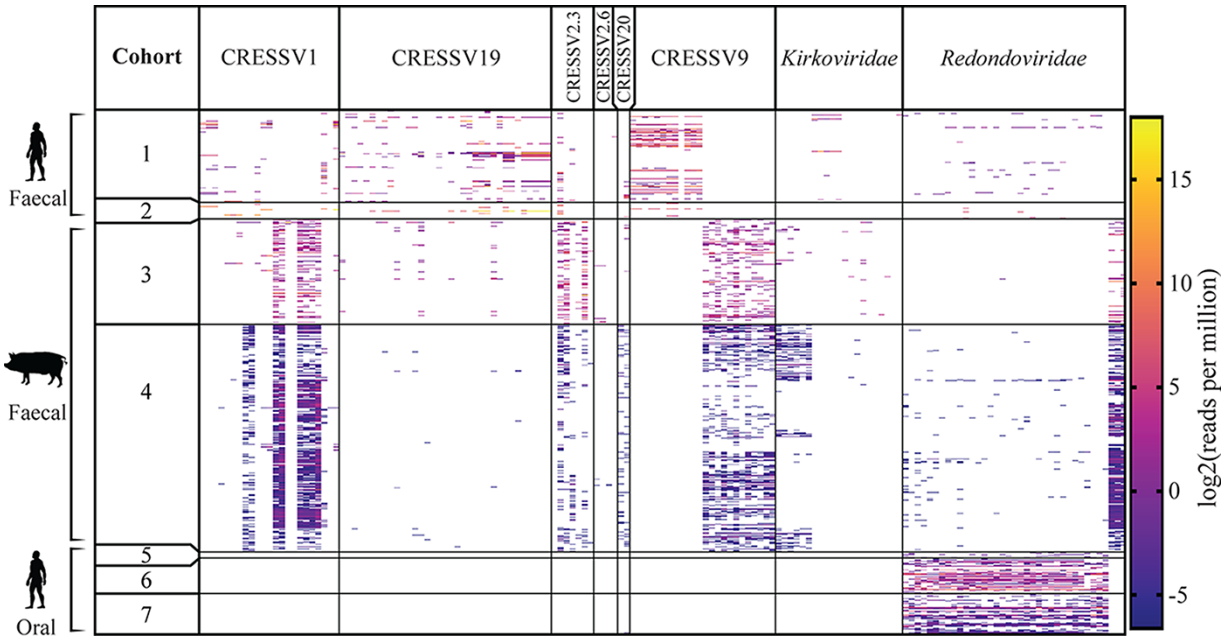


**Figure 3.** Distribution of gastrointestinal cressdnaviruses across seven sample cohorts. Colour represents normalised read count. Empty columns (viruses not found in any sample) and rows (samples containing no viruses) were removed prior to plotting. Members of the CRESSV16 lineage were not detected. Taxon silhouettes are from phylopic.org (*Homo sapiens* by T. Michael Keesey, *Sus scrofa* by Steven Traver). Sample cohorts and viral reference genomes used are reported in Supplementary Tables S2 and S3.

human-associated redondoviruses, and the genus *Entamoeba* was the only host candidate identified. Upon statistical evaluation with Pearson's chi-squared tests, we found that the presence of *Entamoeba* was highly positively associated with the presence of redondoviruses in all three oral cohorts (Supplementary Table S5). Specifically, redondovirus prevalence in subsets of samples positive for *Entamoeba* were 73 per cent, 91 per cent, and 91 per cent, versus 0 per cent, 20 per cent, and 22 per cent in subsets where *Entamoeba* was undetected. In these latter samples, we suspect that if virus was found, non-detection of *Entamoeba* most likely constitutes a false negative. We also found that normalised redondovirus loads were strongly positively correlated with *Entamoeba* loads in the three cohorts, with Spearman's rho values between 0.72 and 0.85 ($P < 0.001$, Supplementary Table S6). At this stage, we therefore predicted *Entamoeba* was the host of redondoviruses. *E. gingivalis* is the only known member of this genus residing in the oral cavity, and examination of BLASTn tables confirmed it was the species identified.

In the two cohorts of human stool samples, presence of the gut protist *Blastocystis* was associated positively with the presence of the CRESSV1 lineage (Supplementary Table S5), with 24 per cent and 9 per cent prevalence in protist positive samples, versus 0 per cent and 1 per cent in negative. Further, CRESSV1 virus loads were positively correlated with *Blastocystis* loads (Supplementary Table S6). The same pattern was observed in both pig stool cohorts; however, in these cases, *Entamoeba* was also associated. While this introduced some uncertainty for host prediction, we noted that prevalences of both protists were extremely high in porcine cohorts, with *Blastocystis* at 76 per cent and 100 per cent prevalence, and *Entamoeba* at 61 per cent and >99 per cent. Normalised loads of both protists were also tightly correlated with each other (cohort 1: rho = 0.72, $P < 0.001$, cohort 2: rho = 0.54, $P < 0.001$), probably due to the shared faecal–oral route of infection, and host factors such as age and health. We suspected the association between CRESSV1 and *Entamoeba* could be driven by this underlying correlation, and we therefore predicted *Blastocystis* (*Blastocystis* spp.) was the likeliest host of CRESSV1, since it was identified and found to be associated in all four stool cohorts, human, and porcine.

Presence of the CRESSV19 lineage was highly positively associated with the presence of *Endolimax* in both human cohorts, and likewise their normalised loads were significantly positively correlated (Supplementary Tables S5 and S6). Importantly, this result was mirrored in both porcine cohorts. Additional associations were found in one of the porcine cohorts, but not both, leading us to predict *E. nana* was the most likely host of CRESSV19.

The result for the kirkoviruses was complex. In both human stool cohorts, presence of kirkoviruses was highly positively associated with the presence of *Dientamoeba*, and likewise their normalised loads were strongly positively correlated. We were therefore surprised when no parabasalid taxa were identified as kirkovirus host candidates in either porcine cohort. Instead, both porcine cohorts showed positive associations to the non-parabasalid genus *Iodamoeba*, in both presence and normalised load. Because evidence from recombination had suggested kirkoviruses at least partly overlapped in host range, we surmised one of the relationships might be incidental. To explore this, we first tested the statistical association between kirkoviruses and *Iodamoeba* in human cohorts, but found none. In the opposite direction, while *Dientamoeba* has been reported in pigs (Cacciò et al. 2012), we found no *Dientamoeba* reads in either porcine cohort. We therefore looked for the presence of other parabasalid taxa. Interestingly, porcine samples highly positive for kirkoviruses did contain a diverse community of parabasalids at high prevalence, including *Trichomitus, Tetratrichomonas, Hypotrichomonas, Trichomonas*, and *Tritrichomonas*. Taken together, at least one parabasalid genus was detected in 10 of 11 samples highly positive for kirkoviruses in cohort 1 and 61 of 62 samples in cohort 2. Since we had previously assumed viruses infected a single genus per cohort type, our host candidate discovery approach would have missed a broader host range. Upon statistical testing, we found significant positive associations between several of the parabasalid genera and kirkoviruses (Supplementary Tables S5 and S6). Despite the lack of clarity in porcine cohorts, due to the strong signal from *Dientamoeba* in human cohorts, we tentatively predicted parabasalids serve as the hosts of kirkoviruses, specifically *D. fragilis* in humans.

Our analyses of CRESSV2.3, CRESSV9, CRESSV20, and pig-associated redondoviruses did not result in host prediction. In the first case, no candidate host taxon was linked to virus presence in human cohorts, although *Iodamoeba* was associated in both porcine cohorts. Testing this genus in human cohorts found no association. Given the high prevalence of parasite infection in porcine cohorts we regarded this as insufficient evidence to predict a host. For both CRESSV9 and CRESSV20, no taxon was identified to be consistently associated with viruses across human and porcine cohorts. In the case of pig-associated redondoviruses, a large set of genera were associated to virus presence in pig stool cohort 1, two of which were also associated in cohort 2 (*Balantioides* and *Balantidium*). Due to the previously mentioned complication of high protist prevalence in porcine samples, we did not make a host prediction.

## Confirmation of host–virus relationships

Our computational workflow predicted protist hosts for four viral lineages: *E. gingivalis* for human-associated redondoviruses, *Blastocystis* spp. for CRESSV1, *E. nana* for CRESSV19, and diverse parabasalid genera for kirkoviruses (specifically *D. fragilis* in humans, and a range of genera in pigs). To independently assess the inferred host–virus relationships, we looked for related EVEs in available protist genome assemblies. No assembly was available for *E. gingivalis*, *E. nana*, or *D. fragilis*, but we included close relatives, and ten *Blastocystis* spp. assemblies (Supplementary Table S7). Notably, four Rep-like EVEs were previously identified in *Blastocystis* spp. (Liu et al. 2011). Our analysis identified thirty-eight cressdnavirus-like EVEs in *Blastocystis* spp., including redetection of the original four. EVEs were distributed across six assemblies from *Blastocystis* spp. subtypes 1, 2, 6, 7, 8, and 9. To confirm their presence in the genome as opposed to assembly contamination, we carried out PCR targeting a subset of six EVEs, using DNA extracted from axenic *Blastocystis* spp. cultures of subtypes 1, 2, 7, and 8. In each case, we could amplify products of the correct size, and two were confirmed by Sanger sequencing. Of the four assemblies in which no EVE was identified, two belonged to subtype 3 and two to subtype 4. Among the thirty-eight EVEs, thirty-seven were Rep-like and one was Cap-like. Clustering of the Rep-like sequences alongside the four of Liu et al. (2011) revealed two distinct clusters and one singleton (Fig. 4A). Cluster 1 included twenty-seven EVEs plus the four previously identified, while cluster 2 contained only newly identified sequences. Phylogenetic analysis confirmed that cluster 2 EVEs belonged to the CRESSV1 virus lineage, validating the prediction that CRESSV1 members infect *Blastocystis* spp. (Fig. 4B, Supplementary Fig. S4A).

Among parabasalids, 145 EVEs were identified in genome assemblies of *Histomonas meleagridis* and 172 were identified in one *Tritrichomonas foetus* assembly. Of the *H. meleagridis* EVEs

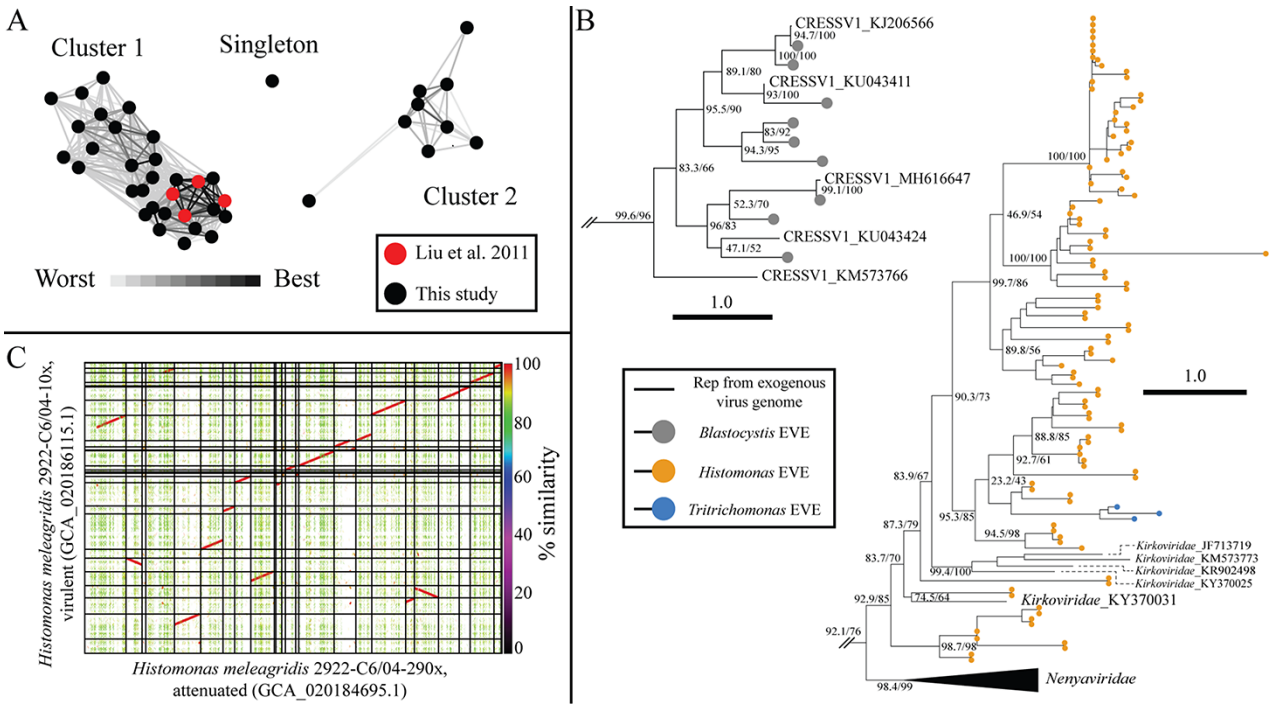


**Figure 4.** EVEs in protist genomes support host inferences. (A) Clustered Rep-like EVEs from *Blastocystis* spp. assemblies. Connections represent significant BLASTp alignments between EVEs, with shade corresponding to level of significance (maximum/worst e-value = 1e-10). Four EVEs identified by Liu et al. (2011) were clustered alongside all thirty-seven Rep-like EVEs detected here. (B) Regions of interest from a phylogeny of Rep-like EVEs and representatives of cressdnavirus lineages (see also Supplementary Fig. S4). Scale bar represents amino acid substitutions per site. (C) Nucmer alignment dotplot between EVE-containing scaffolds from two *Histomonas meleagridis* genome assemblies. Colour denotes alignment percentage similarity. For the list of aligned scaffolds, see Supplementary Table S8.

104 were Rep-like and forty-one were Cap-like, while *T. foetus* EVEs were all Rep-like. Phylogenetic analysis of Rep-like EVEs revealed 102 *H. meleagridis* sequences and three *T. foetus* sequences belonged to the *Kirkoviridae* (Fig. 4B, Supplementary Fig. S4B). This confirms the prediction that kirkoviruses infect parabasalids, although specific validation for *D. fragilis* is still desirable. Notably, the two *H. meleagridis* assemblies were generated from the same strain, one from a virulent form and the other from an attenuated form. Both were originally cultured from a single micro-manipulated cell, with separate passaging for ten or 290 generations, respectively (Palmieri et al. 2021). We thus predicted that scaffolds containing true EVEs would be homologous between such closely related assemblies. Contrastingly, if the sequences actually derived from assembly contamination and were not shared, we would expect dispersal throughout each assembly, and scaffolds would appear mostly non-homologous. We carried out all-vs.-all alignment between EVE-containing scaffolds from the assemblies, twenty-five for GCA_020184695.1 and twenty-nine for GCA_020186115.1 (Supplementary Table S8). The vast majority of scaffolds were clearly homologous, in line with the expectation for true EVEs (Fig. 4C). Notably, these assemblies were built using Oxford Nanopore Technologies long reads in combination with high accuracy Illumina reads, an approach recognised to result in low misassembly rates and high accuracy assemblies (Wick et al. 2017).

Finally, we assessed our prediction that redondoviruses infect *E. gingivalis*. With no host genome assembly available, we ran a case–control screening experiment on DNA extracted from oral plaques of human subjects with periodontitis ($N = 48$), thirty-one with known *E. gingivalis* infection and seventeen tested negative. Samples were screened using qPCR assays for redondoviruses, *E. gingivalis*, and *Trichomonas tenax*. *T. tenax* was included because like *E. gingivalis*, it is a protist associated with human periodontitis (Marty et al. 2017; Benabdelkader et al. 2019), and thus represents an appropriate negative control that should have no association to redondoviruses. We found that qPCR detections of redondoviruses and *E. gingivalis* were highly positively associated with each other (Pearson's chi-squared test: $\chi^2 = 36.71$, $P < 0.001$), while results of redondoviruses and *T. tenax* had no association ($\chi^2 = 0.08$, $P = 0.771$). Using linear regression of Ct values, we additionally found that redondovirus loads were positively correlated with *E. gingivalis* loads ($R^2 = 0.24$, $P = 0.013$, Supplementary Fig. S5), but not with *T. tenax* loads ($R^2 = 0.01$, $P = 0.762$). These results lead us to infer that redondoviruses infect *E. gingivalis*, since they are strongly, consistently, and specifically associated.

## Discussion

Metagenomics has massively expanded known viral diversity. In recognition of the insurmountable task of characterising 'metagenomic species' using traditional laboratory techniques, official taxonomy can now be applied to virus sequence data, rather than characterised isolates alone (Simmonds et al. 2017). In the metagenomic age, host determination is a comparably large and complex task using traditional techniques, with swathes of eukaryotic and prokaryotic taxa intractable to isolation in culture, which also complicates genome sequencing. Here, we developed a metagenomic analysis approach for host prediction that does not rely on a culture system nor a host genome assembly, improving on our previous method (Kinsella et al. 2020). We applied it to metagenomic sequencing datasets containing seven lineages of gastrointestinal cressdnaviruses, several of which have been linked to human and

animal diseases. Host predictions were made for four lineages: human-associated redondoviruses with *E. gingivalis*, kirkoviruses with diverse parabasalid taxa including *D. fragilis* in humans, the CRESSV1 lineage (i.e. pecoviruses) with *Blastocystis* spp., and the CRESSV19 lineage (i.e. hudisaviruses) with *E. nana*. Two of the four predictions (kirkoviruses and lineage CRESSV1) were independently confirmed using EVE evidence, as host genome assemblies were available. For a third prediction (redondoviruses), a case–control experiment was used instead. Our study therefore represents a powerful approach to host identification in the metagenomic age, applicable to any poorly understood virus group found in metagenomic datasets.

Analysis of host presence at the genus level mostly resulted in identification of a single species shared across virus positive samples, yet for kirkovirus hosts in pig stool this resolution was too specific, and was resolved by expanding the taxonomic rank to the Parabasalia. This highlights a complication with utilising taxonomy; equivalent ranks may capture different levels of genetic diversity, and higher ranks may capture the same diversity as lower ones. Illustrating this, the gut-resident amoeba *E. dispar* and *E. histolytica* are closer relatives by rRNA identity than many *Blastocystis* spp. subtypes, and while the former are considered different species, the latter are not (Stensvold et al. 2007). A possible solution for our purpose would be approaching host identity analogously to prokaryotic operational taxonomic units, which apply precise divergence rules to determine taxonomic clusters. Furthermore, while it is broadly true that more closely related viruses are more likely to share hosts, there is no arbitrary genetic divergence cutoff in nature where host switches occur. Purely unsupervised approaches cannot easily address this, and we suggest that the best current solution involves both automated prediction, and expert assessment.

Our findings resolve the possible roles gastrointestinal cressdnaviruses play in human and animal health. Discovered in 2019, the family *Redondoviridae* was found to be strongly associated with human periodontitis and had an observational link to critical illness, but infection of humans has not been demonstrated (Abbas et al. 2019; Zhang et al. 2021). Our finding that the human oral protist *E. gingivalis* is the host of redondoviruses explains their statistical association to periodontitis, since *E. gingivalis* is also strongly linked to gum disease, possibly causally (Bao et al. 2020, 2021; Badri et al. 2021). It implies that redondoviruses do not cause periodontal disease themselves, although it is unknown if they are commensals, or actively modulate host virulence. Some viruses can cause reduced virulence in their hosts, for example, the genomovirus Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1), which severely impacts its phytopathogenic fungal host *Sclerotinia sclerotiorum*, and may represent a potential biocontrol agent (Yu et al. 2010). Whether redondoviruses represent beneficial (or even potentially therapeutic) viruses remains to be explored. Detection of redondoviruses in respiratory samples (from critically ill patients and others) can be explained either by contamination of samples with oral secretions containing shed virus, or by displacement of oral microbiota and secretions to the lung, a particular problem in critical illness and intubation (Scannapieco 1999; Munro and Grap 2004; Blot, Vandijck, and Labeau 2008). Further, we suggest that the relatively rare gut detections of human-associated redondoviruses must represent swallowed virions rather than a site of viral replication. Notably, the *Redondoviridae* are related to the *Naryaviridae*, a family previously found to infect gut-resident species of *Entamoeba* (Kinsella et al. 2020), adding phylogenetic support to the host inference.

We found that lineages CRESSV1 and CRESSV19 also infected protists (*Blastocystis* spp. and *E. nana*, respectively). Both viral lineages have been observed in cases of human diarrhoeal disease (Phan et al. 2016; Altan et al. 2017; Ramos et al. 2021); however, their role was previously ambiguous. We suggest the viruses do not directly influence human disease, but instead indicate underlying protist infection. Both protists have been linked to diarrhoeal disease previously, yet despite millions of annual infections their pathogenicity remains controversial (Scanlan et al. 2014; Poulsen and Stensvold 2016). Similarly, the finding that kirkoviruses infect parabasalid genera has relevance to both human and veterinary health. Kirkoviruses have been identified in dead and diseased livestock on multiple occasions (Li et al. 2015; Guo et al. 2018; Xie et al. 2020), and have also been found in stools of both humans and pigs (Shan et al. 2011; Zhao et al. 2017). While their impact on health remains unmeasured, any such influence must be via biological modulation of their parasite hosts, and our findings provide the basis for answering this. While intriguing, the role of parabasalid infection in previously reported cases of equine disease and death cannot be determined here.

Our study improves the understanding of cressdnavirus ecology. Five cressdnavirus families were already known to infect eukaryotes including plants, vertebrates, algae, and fungi, and three were found to infect protists (Kinsella et al. 2020). Our findings add *Redondoviridae*, CRESSV1, CRESSV19, and *Kirkoviridae* to the latter group, meaning the majority of known cressdnavirus–eukaryote relationships now involve protists. We expect this reflects a broader pattern for the many undetermined relationships remaining.

## Materials and methods

### Cressdnavirus lineage inclusion

A database of cressdnavirus Rep sequences was compiled, containing classified and unclassified lineages. This was aligned to the GenBank nr database (April 2021) using BLASTp (Camacho et al. 2009), and non-redundant cressdnavirus hits were incorporated into the query. This process was iterated two further times, achieving a comprehensive set of 15,815 unique cressdnavirus Reps. Of these, 2,461 remained after clustering with CD-HIT v4.7 (Fu et al. 2012) at 70 per cent global amino acid identity. Reps belonging to the *Arfiviricetes* and *Repensiviricetes* classes were separately aligned using the MUSCLE v5.0.1278 super5 algorithm (Edgar 2021), with -perturb set from 0 to 4 to generate five versions. Best-fit amino acid substitution models were assessed to be VT + G4 + F for all alignments using ModelTest-NG v0.1.6 (Darriba et al. 2020). Maximum likelihood phylogenetic analysis was performed using IQ-TREE v2.1.4-beta (Minh et al. 2020), with settings –ninit 200 -bnni –allnni -B 1000 -alrt 1000. Trees were examined for consistency, and one was annotated per class (that with the highest likelihood score). Unclassified lineages were annotated if the cluster had an UFBoot score ≥85 and at least nine sequences (mean 31 and median 16). Isolation source and host records of annotated sequences were downloaded using Entrez Direct tools (Kans 2013), and used to determine which lineages would be included as 'gastrointestinal cressdnaviruses' (Supplementary Table S1). Strict criteria were not applied, but in practice inclusion required ≥70 per cent of source annotations to be gastrointestinal tract, stool, or wastewater. In the case of human-associated redondoviruses, found predominately in the human oral cavity, respiratory sources were accepted because we considered it plausible they were seeded or contaminated by oral secretions.

## Viral recombination analyses

All available complete genome assemblies from gastrointestinal cressdnavirus lineages were rotated with MARS (Ayad and Pissis 2017) to ensure concordant start positions. Rotated sequences were aligned using MAFFT v7.487 (Katoh, Rozewicki, and Yamada 2017) with automatic settings, and recombination events were analysed using RDP4 v4.101 (Martin et al. 2015). RDP4 was also used to display phylogenetic compatibility and breakpoint pair distribution for the *Redondoviridae*. To construct tanglegrams for each lineage, Rep and Cap proteins were separately aligned using MAFFT v7.487 with automatic settings, and phylogenetic analysis was done using IQ-TREE v1.6.11. Treefiles were loaded into Dendroscope v.3.7.2 (Huson and Scornavacca 2012), rooted at the midpoint, and analysed with the tanglegram algorithm.

## Cressdnavirus distribution across gastrointestinal tract samples

Publically available metagenomic datasets from 1,124 gastrointestinal tract samples belonging to seven cohorts were downloaded (Supplementary Table S2). BWA MEM v0.7.17-r1188 (Li 2013) was used to map reads to 241 gastrointestinal cressdnavirus genomes (Supplementary Table S3). SAM files were processed using the PathoID module of PathoScope v2.0.7 (Hong et al. 2014). False positive mappings were removed by realigning filtered reads to the same genome database using BLASTn with settings -word_size 11 -gapopen 5 -gapextend 2 -penalty -3 -reward 2 -dust yes, and then removing unaligned reads and any with alignment length <40 from the original SAM file. Where original samples were accessible (human stool cohort 1), samples suspected of being false positive due to proximity to a highly positive sample were also curated with PCR (Supplementary Table S9). A matrix of viral distribution covering all cohorts was generated, with empty rows and columns removed. Counts were normalised to reads per million, log2 transformed, and visualised as a heatmap in GraphPad Prism v9.0 (GraphPad Software, San Diego, CA, USA, www.graphpad.com).

## Classification of eukaryotic content in gastrointestinal tract samples

Reads from all 1,124 gastrointestinal samples were mapped to the combined SILVA 138.1 SSU and LSU NR99 databases (Quast et al. 2013). SAM files were processed with PathoScope as above before being filtered to remove bacterial and archaeal hits. Eukaryotic reads were realigned to the GenBank v5 nucleotide database (February 2021) using BLASTn and alignments were filtered with quality cutoffs according to the library preparation method and read length. Specifically, Illumina reads of 100 bp required 100 per cent identity for ≥50 bp, while those ≥150 bp read length required 100 per cent identity for ≥100 bp. VIDISCA IonTorrent reads (Kinsella, Deijs, and van der Hoek 2019) required ≥98 per cent identity for ≥100 bp to allow for possible homopolymer errors. Filtered outputs were processed using Linux command line tools to count occurrences of any specific taxon. Clinically validated qPCRs for *Blastocystis* spp. and *D. fragilis* were run on any sample previously tested for viruses by PCR (Supplementary Table S9), and count tables were updated accordingly.

## Host prediction

Initial host prediction was done on the six of seven cohorts with Illumina deep sequencing data available (Supplementary Table S2). For each viral lineage, samples considered 'highly positive' were selected per cohort. To accommodate variation between different biological lineages and cohorts, we did not apply identical cutoffs, instead treating samples with normalised viral read counts (reads per million) above the inclusive lower quartile value as highly positive. Eukaryotic NCBI taxonomy ID numbers were extracted from the BLASTn tables of these samples, and converted into non-redundant lists of genera using Linux and Entrez Direct tools. Prevalent eukaryotes in highly virus positive samples were then identified using Linux command line tools. Genera normally resident in the gastrointestinal tract were retained, while transient taxa or otherwise implausible identifications were not. We did not apply strict percentage prevalence cutoffs for inclusion as a host candidate, although the lowest was 87.5 per cent (a genus detected in 7 of 8 highly virus positive samples). Next, we tested statistical associations between viruses and respective host candidates across all samples in each separate cohort. Two tests were used; Pearson's chi-squared tests were used to determine if an association existed between presence of a host candidate and a respective cressdnavirus lineage (presence scored 1 and absence scored 0), while Spearman's rank correlation tests were used to determine any correlation between normalised loads of a host candidate and a cressdnavirus lineage. Genera with significant associations to a viral lineage were tested across all cohorts of the same sample type to assess reproducibility. For the main workflow code, see: https://github.com/CormacKinsella/Metagenomic-virus-host-prediction.

## Endogenous viral element analysis

Selected eukaryotic genome assemblies (Supplementary Table S7) were downloaded and searched for Rep and Cap EVEs using tBLASTn (e-value threshold of 1e-5) and a query including 2,923 Rep and 2,122 Cap sequences. Alignment regions were converted to BED format with ascending coordinate ranges, and overlapping features were merged using BEDTools v2.27.1 (Quinlan and Hall 2010). Features were extracted as FASTA sequences, and open reading frames (ORFs) were predicted and translated using EMBOSS v6.3.1 getorf (Rice, Longden, and Bleasby 2000), with settings -minsize 120 -find 0. Virus-like sequences were separated from others using UBLAST v10 (Edgar 2010) and the same query database as above. Filtered candidate EVEs were then aligned to the GenBank nr database with BLASTp, and outputs were inspected to remove false positives. Sequences were clustered using CLANS (Frickey and Lupas 2004). To assess the phylogenetic affiliations of Rep-like EVE sequences, they were aligned alongside five representatives of each cressdnavirus lineage using MAFFT v7.487 E-INS-i, and analysed with IQ-TREE v1.6.11. Based on the results, alignment and phylogenetic analysis was done including all exogenous and endogenous members of the *Kirkoviridae*, using nenyaviruses as an outgroup, and the same for CRESSV1, using vilyaviruses as an outgroup. To confirm *Blastocystis* spp. EVEs were truly found inside genomes and were not assembly contaminants, we extracted genomic DNA from *Blastocystis* spp. axenic cultures belonging to subtypes 1, 2, 7, and 8 using the Boom method (Boom et al. 1990). We then designed and ran PCR assays on extracted DNA to amplify six selected EVEs, and attempted Sanger sequencing of products. To confirm *H. meleagridis* EVEs were genuine, we instead used a computational approach. We carried out all-vs.-all alignment of EVE-containing scaffolds from the two source genome assemblies (built from combined long and short read technologies) using nucmer –maxmatch –nosimplify, within MUMmer v4.0.0rc1 (Marçais et al. 2018). The delta file was then processed using mummerplot.

### Human oral plaque qPCR

Subgingival plaques were collected with curettes from inflamed periodontal pockets of patients with clinically diagnosed periodontitis, at the Department of Periodontology, Oral Medicine and Oral Surgery, Charité—Universitätsmedizin Berlin. Plaque was directly transferred into lysis buffer and DNA was extracted by the phenol/chloroform method. Samples were PCR screened using *E. gingivalis* specific primers (Bonner et al. 2014), and the human gene *ACTB* (beta-actin) was also amplified as a DNA isolation control (Supplementary Table S9). For this study, forty-eight DNA extractions with sufficient residual material were selected, comprising thirty-one *E. gingivalis* positive and seventeen negative samples. Three TAMRA qPCR assays targeting *Redondoviridae, E. gingivalis*, and *T. tenax* were designed (Supplementary Table S9), and tenfold dilutions of each target were used to construct standard curves and determine cycle threshold limits (Ct values ≥37 were considered negative). All forty-eight samples were screened once for the three targets alongside standards and negative controls. Association between test outcomes (positive scored 1 and negative scored 0) was assessed using Pearson's chi-squared test, and correlation between Ct values was explored using linear regression.

### Ethical approval

Work with clinical samples from human subjects was approved by a vote of the local ethics committee (Campus Charité Mitte, application number EA1/169/20).

## Data availability

All sequence datasets and genome assemblies utilised here are available in public databases; see Supplementary Tables S2, S3, and S7. For workflow code, see: https://github.com/CormacKinsella/Metagenomic-virus-host-prediction.

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

## Acknowledgements

Computational work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

## Funding

This work was supported by a grant from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie agreement No. 721367 (HONOURs), awarded to Lia van der Hoek.

**Conflict of interest:** None declared.

## References

Abbas, A. A. et al. (2019) '*Redondoviridae*, a Family of Small, Circular DNA Viruses of the Human Oro-respiratory Tract that are Associated with Periodontitis and Critical Illness', *Cell Host & Microbe*, 25: 719–29.

Ahlgren, N. A. et al. (2017) 'Alignment-free $D_2^*$ Oligonucleotide Frequency Dissimilarity Measure Improves Prediction of Hosts from Metagenomically-derived Viral Sequences', *Nucleic Acids Research*, 45: 39–53.

Altan, E. et al. (2017) 'Small Circular Rep-encoding Single-stranded DNA Genomes in Peruvian Diarrhea Virome', *Genome Announcements*, 5: e00822–17.

Ayad, L. A. K., and Pissis, S. P. (2017) 'MARS: Improving Multiple Circular Sequence Alignment Using Refined Sequences', *BMC Genomics*, 18: 1–10.

Babayan, S. A., Orton, R. J., and Streicker, D. G. (2018) 'Predicting Reservoir Hosts and Arthropod Vectors from Evolutionary Signatures in RNA Virus Genomes', *Science*, 362: 577–80.

Badri, M. et al. (2021) 'Current Global Status and the Epidemiology of *Entamoeba Gingivalis* in Humans: A Systematic Review and Meta-analysis', *Acta Parasitologica*, 66: 1102–13.

Bao, X. et al. (2021) '*Entamoeba Gingivalis* Exerts Severe Pathogenic Effects on the Oral Mucosa', *Journal of Dental Research*, 100: 771–6.

——— et al. (2020) '*Entamoeba Gingivalis* Causes Oral Inflammation and Tissue Destruction', *Journal of Dental Research*, 99: 561–7.

Benabdelkader, S. et al. (2019) 'Specific Clones of *Trichomonas Tenax* are Associated with Periodontitis', *PLOS One*, 14: e0213338.

Bickhart, D. M. et al. (2019) 'Assignment of Virus and Antimicrobial Resistance Genes to Microbial Hosts in a Complex Microbial Community by Combined Long-read Assembly and Proximity Ligation', *Genome Biology*, 20.

Blot, S., Vandijck, D., and Labeau, S. (2008) 'Oral Care of Intubated Patients', *Clinical Pulmonary Medicine*, 15: 153–60.

Bonner, M. et al. (2014) 'Detection of the Amoeba *Entamoeba Gingivalis* in Periodontal Pockets', *Parasite*, 21.

Boom, R. et al. (1990) 'Rapid and Simple Method for Purification of Nucleic Acids', *Journal of Clinical Microbiology*, 28: 495–503.

Cacciò, S. M. et al. (2012) 'Pigs as Natural Hosts of *Dientamoeba fragilis* Genotypes Found in Humans', *Emerging Infectious Diseases*, 18: 838–41.

Camacho, C. et al. (2009) 'BLAST+: Architecture and Applications', *BMC Bioinformatics*, 10: 421.

Darriba, D. et al. (2020) 'ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models', *Molecular Biology and Evolution*, 37: 291–4.

Díez-Villaseñor, C., and Rodriguez-Valera, F. (2019) 'CRISPR Analysis Suggests that Small Circular Single-stranded DNA Smacoviruses Infect *Archaea* Instead of Humans', *Nature Communications*, 10.

Dion, M. B. et al. (2021) 'Streamlining CRISPR Spacer-based Bacterial Host Predictions to Decipher the Viral Dark Matter', *Nucleic Acids Research*, 49: 3127–38.

Dolja, V. V., and Koonin, E. V. (2018) 'Metagenomics Reshapes the Concepts of RNA Virus Evolution by Revealing Extensive Horizontal Virus Transfer', *Virus Research*, 244: 36–52.

Duffy, S., Burch, C. L., and Turner, P. E. (2007) 'Evolution of Host Specificity Drives Reproductive Isolation among RNA Viruses', *Evolution*, 61: 2614–22.

Edgar, R. C. (2010) 'Search and Clustering Orders of Magnitude Faster than BLAST', *Bioinformatics*, 26: 2460–1.

——— (2021) 'MUSCLE V5 Enables Improved Estimates of Phylogenetic Tree Confidence by Ensemble Bootstrapping', *BioRxiv*.

——— et al. (2022) 'Petabase-scale Sequence Alignment Catalyses Viral Discovery', *Nature*, 602: 142–7.

Ellis, J. et al. (1998) 'Isolation of Circovirus from Lesions of Pigs with Postweaning Multisystemic Wasting Syndrome', *The Canadian Veterinary Journal*, 39: 44–51.

Eng, C. L. P., Tong, J. C., and Tan, T. W. (2014) 'Predicting Host Tropism of Influenza A Virus Proteins Using Random Forest', *BMC Medical Genomics*, 7.

Frickey, T., and Lupas, A. (2004) 'CLANS: A Java Application for Visualizing Protein Families Based on Pairwise Similarity', *Bioinformatics*, 20: 3702–4.

Fu, L. et al. (2012) 'CD-HIT: Accelerated for Clustering the Next-generation Sequencing Data', *Bioinformatics*, 28: 3150–2.

Greninger, A. L. (2018) 'A Decade of RNA Virus Metagenomics Is (Not) Enough', *Virus Research*, 244: 218–29.

Guo, Z. et al. (2018) 'Identification and Genomic Characterization of a Novel CRESS DNA Virus from a Calf with Severe Hemorrhagic Enteritis in China', *Virus Research*, 255: 141–6.

Harding, R. M. et al. (1993) 'Nucleotide Sequence of One Component of the Banana Bunchy Top Virus Genome Contains a Putative Replicase Gene', *Journal of General Virology*, 74: 323–8.

Hong, C. et al. (2014) 'PathoScope 2.0: A Complete Computational Framework for Strain Identification in Environmental or Clinical Sequencing Samples', *Microbiome*, 2.

Huson, D. H., and Scornavacca, C. (2012) 'Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks', *Systematic Biology*, 61: 1061–7.

Ignacio-Espinoza, J. C. et al. (2020) 'Ribosome-linked mRNA-rRNA Chimeras Reveal Active Novel Virus Host Associations', *BioRxiv*.

Kans, J. (2013). Entrez Direct: E-utilities on the Unix Command Line. <https://www.ncbi.nlm.nih.gov/books/NBK179288/> accessed 19 Oct 2021.

Kapoor, A. et al. (2010) 'Use of Nucleotide Composition Analysis to Infer Hosts for Three Novel Picorna-like Viruses', *Journal of Virology*, 84: 10322–8.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2017) 'MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization', *Briefings in Bioinformatics*, 20: 1160–6.

Kazlauskas, D., Varsani, A., and Krupovic, M. (2018) 'Pervasive Chimerism in the Replication-associated Proteins of Uncultured Single-stranded DNA Viruses', *Viruses*, 10.

Kinsella, C. M. et al. (2020) '*Entamoeba* and *Giardia* Parasites Implicated as Hosts of CRESS Viruses', *Nature Communications*, 11: 1–10.

Kinsella, C. M., Deijs, M., and van der Hoek, L. (2019) 'Enhanced Bioinformatic Profiling of VIDISCA Libraries for Virus Detection and Discovery', *Virus Research*, 263: 21–6.

Kirk, M. D. et al. (2015) 'World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis', *PLOS Medicine*, 12: e1001921.

Krupovic, M. et al. (2020) 'Cressdnaviricota: A Virus Phylum Unifying Seven Families of Rep-Encoding Viruses with Single-Stranded, Circular DNA Genomes', *Journal of Virology*, 94.

Lefeuvre, P. et al. (2007) 'Avoidance of Protein Fold Disruption in Natural Virus Recombinants', *PLOS Pathogens*, 3: e181.

——— et al. (2009) 'Widely Conserved Recombination Patterns among Single-stranded DNA Viruses', *Journal of Virology*, 83: 2697–707.

Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. ArXiv:1303.3997v1 [q-Bio.GN].

Li, L. et al. (2015) 'Exploring the Virome of Diseased Horses', *The Journal of General Virology*, 96: 2721–33.

Liu, D. et al. (2019) 'Predicting Virus-host Association by Kernelized Logistic Matrix Factorization and Similarity Network Fusion', *BMC Bioinformatics*, 20.

Liu, H. et al. (2011) 'Widespread Horizontal Gene Transfer from Circular Single-stranded DNA Viruses to Eukaryotic Genomes', *BMC Evolutionary Biology*, 11.

Magee, C. J. (1927) 'Investigation on the Bunchy Top Disease of the Banana', *Bulletin Council for Scientific and Industrial Research*, 30.

Marçais, G. et al. (2018) 'MUMmer4: A Fast and Versatile Genome Alignment System', *PLOS Computational Biology*, 14: e1005944.

Martin, D. P. et al. (2015) 'RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes', *Virus Evolution*, 1.

Marty, M. et al. (2017) '*Trichomonas tenax* and Periodontal Diseases: A Concise Review', *Parasitology*, 144: 1417–25.

Minh, B. Q. et al. (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37: 1530–4.

Munro, C. L., and Grap, M. J. (2004) 'Oral Health and Care in the Intensive Care Unit: State of the Science', *American Journal of Critical Care*, 13: 25–34.

Nagasaki, K. et al. (2005) 'Previously Unknown Virus Infects Marine Diatom', *Applied and Environmental Microbiology*, 71: 3528–35.

Palmieri, N. et al. (2021) 'Complete Genomes of the Eukaryotic Poultry Parasite *Histomonas meleagridis*: Linking Sequence Analysis with Virulence/Attenuation', *BMC Genomics*, 22.

Phan, T. G. et al. (2016) 'The Fecal Virome of South and Central American Children with Diarrhea Includes Small Circular DNA Viral Genomes of Unknown Origin', *Archives of Virology*, 161: 959–66.

Poulsen, C. S., and Stensvold, C. R. (2016) 'Systematic Review on *Endolimax nana*: A Less Well Studied Intestinal Ameba', *Tropical Parasitology*, 6: 8–29.

Quast, C. et al. (2013) 'The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-based Tools', *Nucleic Acids Research*, 41: D590–6.

Quinlan, A. R., and Hall, I. M. (2010) 'BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features', *Bioinformatics*, 26: 841–2.

Ramos, E. D. S. F. et al. (2021) 'Composition of Eukaryotic Viruses and Bacteriophages in Individuals with Acute Gastroenteritis', *Viruses*, 13.

Rice, P., Longden, L., and Bleasby, A. (2000) 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends in Genetics*, 16: 276–7.

Ritchie, B. W. et al. (1989) 'Characterization of a New Virus from Cockatoos with Psittacine Beak and Feather Disease', *Virology*, 171: 83–8.

Scanlan, P. D. et al. (2014) 'The Microbial Eukaryote *Blastocystis* Is a Prevalent and Diverse Member of the Healthy Human Gut Microbiota', *FEMS Microbiology Ecology*, 90: 326–30.

Scannapieco, F. A. (1999) 'Role of Oral Bacteria in Respiratory Infection', *Journal of Periodontology*, 70: 793–802.

Shan, T. et al. (2011) 'The Fecal Virome of Pigs on a High-density Farm', *Journal of Virology*, 85: 11697–708.

Shi, M. et al. (2016) 'Redefining the Invertebrate RNA Virosphere', *Nature*, 540: 539–43.

Simmonds, P. et al. (2017) 'Virus Taxonomy in the Age of Metagenomics', *Nature Reviews. Microbiology*, 15: 161–8.

Stensvold, C. R. et al. (2007) 'Terminology for *Blastocystis* Subtypes – A Consensus', *Trends in Parasitology*, 23: 93–6.

Tam, C. C. et al. (2012) 'Changes in Causes of Acute Gastroenteritis in the United Kingdom over 15 Years: Microbiologic Findings from 2 Prospective, Population-based Studies of Infectious Intestinal Disease', *Clinical Infectious Diseases*, 54: 1275–86.

Thumbi, S. M. et al. (2015) 'Linking Human Health and Livestock Health: A "One-health" Platform for Integrated Analysis of Human Health, Livestock Health, and Economic Welfare in Livestock Dependent Communities', *PLOS One*, 10: e0120761.

Tisza, M. J. et al. (2020) 'Discovery of Several Thousand Highly Diverse Circular DNA Viruses', *Elife*, 9: e51971.

Varma, A., and Malathi, V. G. (2003) 'Emerging Geminivirus Problems: A Serious Threat to Crop Production', *Annals of Applied Biology*, 142: 145–64.

Wick, R. R. et al. (2017) 'Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads', *PLOS Computational Biology*, 13: e1005595.

Xie, J. et al. (2020) 'First Detection and Genetic Characterization of a Novel Kirkovirus from a Dead Thoroughbred Mare in Northern Xinjiang, China, in 2018', *Archives of Virology*, 165: 403–6.

Yoon, H. S. et al. (2011) 'Single-cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists', *Science*, 332: 714–7.

Yu, X. et al. (2010) 'A Geminivirus-related DNA Mycovirus that Confers Hypovirulence to a Plant Pathogenic Fungus', *Proceedings of the National Academy of Sciences*, 107: 8387–92.

Zhang, Y. et al. (2021) '*Redondoviridae* and Periodontitis: A Case–Control Study and Identification of Five Novel Redondoviruses from Periodontal Tissues', *Virus Evolution*, 7: 33.

Zhao, G. et al. (2017) 'Intestinal Virome Changes Precede Autoimmunity in Type I Diabetes-susceptible Children', *Proceedings of the National Academy of Sciences*, 114: E6166–75.

Zhao, L., Lavington, E., and Duffy, S. (2021) 'Truly Ubiquitous CRESS DNA Viruses Scattered across the Eukaryotic Tree of Life', *Journal of Evolutionary Biology*, 34: 1901–16.